Arch Sci
DOI 10.1007/s10502-012-9172-7

ORIGINAL PAPER

Author Proof

Provenance and credibility in scientific data repositories

Kathleen Fear · Devan Ray Donaldson

© Springer Science+Business Media B.V. 2012

Abstract Despite a long history of rich theoretical work on provenance, empirical research regarding users' interactions with and judgments based upon provenance information in archives with scientific data is extremely limited. This article focuses on the relationship between provenance and credibility (i.e., trustworthiness and expertise) for scientists. Toward this end, the authors conducted semi-structured interviews with seventeen proteomics researchers who interact with data from ProteomeCommons.org, a large online repository. To analyze the resulting interview data, the authors apply Brian Hilligoss and Soo Young Rieh's empirically tested theoretical framework for user credibility assessment. Findings from this study suggest that together with other information provided in ProteomeCommons.org and subjects' own experiences and prior knowledge, provenance allows users to determine the credibility of datasets. Implications of this study stress the importance of the archival perspective of provenance and archival bond for aiding scientists in their credibility assessments of data housed in scientific data repositories.

Keywords Provenance · Credibility · Scientific data · Metadata

Introduction

Archival scholars have posited that provenance is critical to retrieval (Bearman and Lytle 1985), contextualization (Cook 2001), and assessing the authenticity (Duranti, 2001) of records. The concept of provenance has, however, been primarily developed within the context of traditional archives housing bureaucratic records.

K. Fear (✉) · D. R. Donaldson
School of Information, University of Michigan, 3339D North Quad, 105 S. State Street,
Ann Arbor, MI 48109-1285, USA
e-mail: kfear@umich.edu

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | □ LE | □ TYPESET
MS Code : ARCS185 | ☑ CP | ☑ DISK

Author Proof

More recently, archival scholars have turned their attention to other kinds of records and the archives they comprise. Of particular interest are scientific research data.

Among the broad range of topics covered in *Archival Science*'s special issue on archiving research data, the principle of provenance came up repeatedly. Vardigan and Whiteman (2007) mention provenance as one of the types of information that must be included in Archival Information Packages (AIPs). They state that ICPSR, a major social science data archive, describes some provenance information in associated metadata records for data (including bibliographic information, collection changes, and history) and that this information is accessible to users. Corti (2007) states that research data archive "add value" to data "by providing enhanced resource discovery and richer comprehension about the data and its provenance" (p. 48) and that this enrichment is critical to increasing visibility as well as enabling easier and more effective use of data by researchers and teachers. Studies have not, however, focused on how end users actually interact with provenance information. While provenance has been identified as important information for users, there is little empirical research on how users actually use provenance.

This study seeks to add to discussion concerning archiving research data by conducting empirical research to examine users' interactions with provenance. As a test case, we focus on a specific scientific community—proteomics researchers—of a major scientific data repository, ProteomeCommons.org, and the provenance metadata made available in that repository through the Minimal Information About a Proteomics Experiment (MIAPE) standard, a community-developed metadata framework. Data reuse is critical to the advancement of research in proteomics, and provenance information such as how data were collected, who performed the data collection, how the project was funded, and what has happened to the data during its lifetime is, in turn, critical to enabling reuse. However, when interacting with data, users have access to more than just provenance information; thus, we situate users' interactions with provenance information about data within the larger context of establishing the credibility of a dataset using Hilligoss and Rieh's (2008) framework for credibility assessment. Trustworthiness and expertise are primary indicators of credibility, and our hypothesis was that provenance contributes to user assessment of both these criteria. Our primary research questions are the following: How do users determine credibility (i.e., trustworthiness and expertise) in the context of a data repository? To what extent do provenance metadata facilitate users' credibility assessments by providing information about trustworthiness and expertise?

Literature review

The computer science perspective on provenance has dominated work on scientific data archives. Buneman et al. (2001, p. 316) characterize data provenance as "the origin of a piece of data and the process by which it arrived in a database." This perspective focuses on the manipulability of data: Particularly in the case of reuse, data are often transformed, recombined, or otherwise removed from their original contexts. Because data are so mutable, confirmation of results based on a particular dataset as well as reuse of those data is contingent on an understanding of their

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

provenance. Greenwood et al. (2003) expand upon Buneman et al.'s definition, incorporating the idea of workflow and experimental notes into data provenance. Much of this work on provenance in science has emphasized the utility of provenance information for establishing trustworthiness, typically understood as confidence in the integrity and quality of data (Bertino et al. 2009; Dai et al. 2008). Bose and Frew (2005) suggested a number of benefits that provenance information—or what they term "lineage systems"—makes possible, including the ability to identify the source of faulty or anomalous outputs, saving processing "recipes" and being able to rerun processing sequences, and comparing the analytical steps involved in producing the data in two or more databases. Simmhan et al. (2005) conducted a comprehensive survey of data provenance techniques for scientific data, finding that while there are a number of systems that include aspects of data provenance, there are still open questions about how best to encode provenance metadata and ensure data trustworthiness.

Studies of provenance metadata for scientific data have taken a heavily technology-focused approach, emphasizing the development of systems for tracking and generating provenance metadata. This is in line with much of the previous work in metadata management in the sciences more generally, which has focused primarily on technical infrastructure, especially the components that are necessary to create lineage systems (Bose and Frew 2005) and tying provenance metadata generation to workflow tools and applications (Bowers et al. 2006; Heinis and Alonso 2008), along with the development of conceptual models for data lineage (Bose 2002). Buneman et al. (2006) proposed a database system that would preserve provenance information when data are copied across databases and integrated into new datasets.

In contrast, the archival perspective on provenance incorporates more than data lineage. There are significant social aspects that come into play, particularly with respect to the relationships not only between individual records but among "the organizations or individuals that created, accumulated and/or maintained and used them in the conduct of personal or corporate activity" (Society of American Archivists 2004, p. 206). Archival provenance emphasizes the functional and structural contexts of records and their evolution over time (Cook 2001). Provenance is also a concern of researchers in digital preservation: Provenance in the context of digital records is referred to as "the record of the chain of custody and change history of a digital object" (Caplan 2009, p. 6), which extends provenance to incorporate the repository and its preservation actions. International standards have been developed for capturing digital provenance postingest as events, for example, when repositories have made changes to their digital objects as a means of preserving intellectual content (e.g., migrating digital objects from an older file format to a newer one) (PREMIS Editorial Committee 2008).

Despite a long history of rich theoretical work on provenance, empirical research regarding users' interactions with and judgments about provenance information in archives generally and with scientific data in particular is extremely limited. Lauriault et al. ~~(2008)~~ found that users want access to "metadata that include lineage information to help them determine if the data are fit for use." Fit-for-use decisions are based on several factors. Lauriault et al. list accuracy, authenticity, and

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

reliability, while other studies emphasize the importance of the identity of the data creator, and particularly their membership in a community of practice (Van House 2002; Zimmerman 2008). As used by Hilligoss and Rieh (2008), trustworthiness and expertise are conceptually parallel to Lauriault et al.'s categories: An information object is trustworthy if "it appears to be reliable, unbiased, and fair" (p. 1469) (if it is accurate) and credible if the source is known to have sufficient and appropriate expertise (if it is authentic and reliable). This suggests that provenance, by facilitating assessments of authenticity, reliability, and accuracy, participates in users' process of determining credibility. To examine how users interact with provenance and other kinds of information during their credibility assessments, we draw on Hilligoss and Rieh's empirically tested credibility assessment framework, which describes users' interaction with credibility indicators at construct, heuristic, and interaction levels.

Construct: what are the characteristics of a credible piece of information? (pp. 1474–1475)

This level encompasses the broad concepts that indicate credibility. The concepts include truthfulness, believability, trustworthiness, and objectivity. Different users may draw on different subsets of these concepts depending on their needs and the kind of information they are looking for, but common to all these concepts is that they are difficult to measure empirically.

Heuristic: what rules of thumb allow for quick judgments about credibility? (pp. 1475–1477)

Because certain constructs are hard to assess directly, information seekers rely on "rules of thumb" about easily observable features of information that indicate the presence of particular constructs. These characteristics can be *media-based* (e.g., a rule of thumb may be that books are more truthful than magazines), *source-related* (e.g., some authors may be more believable than others), *endorsement-based* (e.g., a work that is cited often may be more authoritative than one cited infrequently), or *esthetics-based* (e.g., a well-designed Web site could be more trustworthy than one that is shoddily put together).

Interaction: what particular characteristics of a piece of information speak to its credibility? (pp. 1477–1479)

The construct and heuristic levels help assess credibility at a very general level. The interaction level describes how users assess credibility in the context of using a specific information object. There are three kinds of interactions that can yield information about credibility. First, users can interact with the content itself. They can also interact with peripheral source cues, such as their knowledge of the institution an author is affiliated with and peripheral information object cues, like the amount or completeness of the available metadata.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Methods

This study considers the following research questions: How do users determine credibility (i.e., trustworthiness and expertise) in the context of a data repository? To what extent do provenance metadata facilitate users' credibility assessments by providing information about trustworthiness and expertise? To explore these questions, we conducted seventeen interviews with proteomics researchers to: 1) gather information about how users determine whether a dataset is credible and 2) explore what information users feel is necessary and helpful to them in evaluating data in the context of a data repository.

What is proteomics?

Proteomics is what is known as a postgenome science. It is one of the many new lines of scientific inquiry opened as a result of advances in genome sequencing. Genes produce proteins; the entire complement of proteins produced by a genome is known as the "proteome." In the same way that every organism has a different genome, every organism has a unique proteome. However, unlike an organism's genome, which remains constant throughout its lifetime, the proteome of an organism—or even the proteins present in different cells within a single organism—will change over time as different genes are expressed or inhibited. This makes sequencing a given proteome a somewhat more complex problem than genome sequencing. However, the dynamic nature of the proteome yields important information to researchers. For example, changes to the types and amount of proteins in a cell or organism can correlate with different disease states. By identifying these changes, researchers can isolate biomarkers, which can then be used to diagnose diseases quickly and accurately.

Proteomics researchers use a variety of techniques and instruments, including mass spectrometry and gel electrophoresis, to isolate, sequence, and identify proteins. While many researchers are involved in studies to determine the biological significance of proteins, others are engaged in developing new methods to more accurately sequence and identify proteins, especially those that are present at very low concentrations. Like genomics, proteomics is a high-throughput, data-intensive science, and there is a significant benefit to be had in reusing the massive amounts of data the field produces. In recognition of this fact, there are several databases that collect proteomics data and metadata; further, *Molecular and Cellular Proteomics*, the flagship journal in the field, now requires that any author who submits an article using mass spectrometry data must make that data publicly available (Rodriguez et al. 2010).

Provenance in proteomics

As a field, proteomics has already taken steps toward standardizing metadata practices. The Minimal Information About a Proteomics Experiment (MIAPE) standard provides a set of metadata elements that comprise the minimal amount of information necessary to capture all relevant information about a dataset and the experiment that produced it (Taylor et al. 2007). This standard has been widely

Author Proof
UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

implemented and is used in several major repositories for proteomics data. The standard includes elements that, from our perspective as archivists, we considered provenance information, including the date on which the data were initiated; the name(s) of the person(s) responsible for the creation of the data; information about data transformation techniques used; analysis tools used; and information about data generation, including the location of the raw data, databases queried or specifications of equipment and conditions under which the data were produced. Table 1 shows the provenance elements we identified in the standard. The metadata fields available in ProteomeCommons.org are based on the MIAPE standard, a metadata standard developed by the proteomics community. This framework as implemented by ProteomeCommons.org allows authors to provide extensive metadata if they so choose, but there is no minimum requirement.

ProteomeCommons.org

ProteomeCommons.org is one of the major proteomics data repositories, containing about 11Terabytes of data provided by authors or harvested from other proteomics data systems. This repository, housed at the University of Michigan, provides a data annotation system to researchers, allowing them to supply metadata about the data they submit to Tranche, a repository system that is integrated with ProteomeCommons.org. Users of ProteomeCommons.org include academic researchers, researchers affiliated with non-university-based research centers, and researchers in industry settings. Proteomics is an international field, and researchers from all over the world are represented in the ProteomeCommons.org user community.

The MIAPE standard is implemented in ProteomeCommons.org, and in the current release version of ProteomeCommons.org, some provenance metadata are displayed, including information about who uploaded data and when, his/her name, country, organization/institution, and department. There is also a "Description" section, which is a free-text area that may also contain provenance information, such as the study's funding source or other individuals who worked on the dataset. There are no provenance elements in the MIAPE standard that are not included in ProteomeCommons.org's implementation, but ProtoemeCommons.org also includes additional indicators of a dataset's integrity. At the top of the data download page is

Table 1 Provenance elements identified in the MIAPE metadata standard

Element Name	Description
Date Stamp	The date on which the work described was initiated; given in the standard "YYYY-MM-DD" format (with hyphens).
Responsible person or institutional role	The (stable) primary contact person for this dataset; this could be the experimenter, laboratory head, line manager, etc.
Data transformation techniques	Include algorithms used, preparation or processing techniques, normalization techniques.
Analysis tools	Include software name and version, initial input parameters.
Data generation	Includes location of raw data, databases queried or specifications of equipment and conditions under which data were produced.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

an alphanumeric string, or the Tranche hash, which can be used to verify that data have not changed since they were published. The hash is a checksum calculated based on the content of the dataset plus its license and a small set of metadata and is unique for that dataset. A researcher looking to use the data can download the dataset and recalculate the hash. If the new hash matches what is displayed in ProteomeCommons.org, the researcher can be confident that the data are identical to what were uploaded. Differences in the hash codes would indicate that the data have been corrupted.

Data collection, instruments, and analysis

ProteomeCommons.org had 581 registered users at the time our study began.[1] We excluded users who were on ProteomeCommons.org's development team as well as those who had registered but never uploaded data, resulting in a pool of 191 eligible subjects. We recruited participants for interviews via e-mail. Because our subject base was globally distributed, we conducted phone interviews with users in the United States and Canada ($n = 13$) and sent e-mail versions of the same protocol to users in Europe and Asia ($n = 4$). Every subject, regardless of the primary interview mode, also filled out a demographic survey by e-mail at the end of the interview. We completed 17 total interviews between June 2010 and August 2010.

The interviews consisted of a set of questions (See Appendix 1) relating to the participant's data deposit and reuse behaviors, including assessing a sample publication and the information that paper provided about the data used therein. Participants also completed a rating exercise (See Appendix 2). In the exercise, participants were provided with a list of provenance elements that are included in the MIAPE standard, and they were asked to rate on a five-point Likert scale the confidence they would have in a dataset based only on that information. The demographic survey recorded each subject's level of experience with proteomics research and their current employment, among other information (See Appendix 3). We developed a coding schema based on Hilligoss and Rieh's credibility framework, described above, and coded the interview transcripts using NVivo (Bazeley 2007). Because of the small sample size, no statistical methods were used on the quantitative data.

Findings

Study subjects

The subjects in this study are representative of the larger ProtoemeCommons.org user base: Many are located in the United States, but four work abroad, and while the majority (8) were faculty members, the rest are a mix of postdoctoral fellows/

[1] The actual number of users is likely higher, since users only need to register to upload datasets, not to download them.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

researchers (5), staff scientists (3), and one consultant. They have a range of experience in the field. Most had worked in the proteomics research area for 1–5 years (9) or had 5–10 years of experience (5). Table 2 includes attributes of our subjects.

All of our subjects had experience interacting with ProteomeCommons.org. Except for one individual, our subjects indicated that they themselves are responsible for uploading data (rather than passing the job off to a graduate student, for example). Among our subjects, the most common reason for using ProteomeCommons.org was to fulfill publication requirements that stipulated datasets must be accessible; eight subjects indicated this. Six subjects use ProteomeCommons.org to share data, two use it primarily as a portal for accessing Tranche, the database back end of the system, and one subject uses it to make datasets broadly available. Close to half (7) had also downloaded and used data from ProteomeCommons.org: Four had used an entire dataset, one had used part of a dataset, and two had done both. Two of the ten subjects who said they had not used data from ProteomeCommons.org left clarifying comments on the demographic survey (Appendix 3): One indicated that she did use data, but it was strictly from collaborators, and the other said he had downloaded data, but only as part of the process of reviewing a paper. The number of times our subjects had uploaded data to ProteomeCommons.org ranged from one to 72, with the majority of users (10) classified as "heavy" users (number of uploads >5). Six were light users, only having uploaded once or twice. None of our subjects, however, perceived themselves to be "heavy" users; eight felt that they were "medium" users, and nine thought they were "light" users.

We initially assumed that (1) researchers rely primarily on provenance metadata when initially evaluating a dataset and (2) some elements of provenance metadata

Table 2 Study subjects' demographic attributes

Subjects	Rank	Location
01	Assistant Professor	Europe
02	Assistant Professor	Canada
03	Proteomics Consultant	U. S.
04	Professor	U. S.
05	Staff Scientist	U. S.
06	Post-Doctoral Fellow/Researcher	U. S.
07	Assistant Professor	U. S.
08	Post-Doctoral Fellow/Researcher	U. S.
09	Staff Scientist	U. S.
10	Assistant Research Professor	U. S.
11	Post-Doctoral Fellow/Researcher	U. S.
12	Assistant Professor	U. S.
13	Assistant Research Professor	U. S.
14	Post-Doctoral Fellow/Researcher	U. S.
15	Post-Doctoral Fellow/Researcher	Europe
16	Assistant Professor	Canada
17	Staff Scientist	U. S.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

would be more useful than others. As we conducted interviews and then analyzed the resulting data, however, it became clear that these assumptions did not hold. We found that these users paid attention to provenance but in a somewhat different manner than we expected. Specifically, provenance provided a foundation for evaluating a dataset's trustworthiness and the expertise of the dataset's source. Together with other information provided in ProteomeCommons.org and our subjects' own experience and prior knowledge, provenance allowed users to determine the credibility of datasets. The following sections explore in more detail how users use provenance to establish a dataset's credibility (i.e., data trustworthiness and the expertise of its source) and finally, how provenance fits into the larger picture of the information environment in which credibility assessments occur.

Using provenance an indicator of trustworthiness

Our questions focused largely on subjects' evaluations of the way they and others document their data, both as citations or descriptions in papers and as metadata contributed to ProteomeCommons.org and Tranche. For the most part, researchers expressed satisfaction with what others provided and confidence in their own documentation processes: S16 felt he provided enough information for others to trust and use his data because he'd "never had anybody complain so far." S14 indicated that she provides "as much information as possible in a text format ready to be cited by the re-users." S04 described providing enough information for others to trust and use his data as a personal goal, stating "If we are going to have a public repository and I'm going to go to the lengths to put it there, I certainly would be happy to hear that someone's using the data." For S04, providing appropriate information about the datasets he uploads to ProteomeCommons.org would serve as a means to such an end. In part, simply making data available contributes to its trustworthiness. One subject indicated that publishing or archiving data is a form of endorsement:

> When somebody is willing to stick their neck out and make their data available, I think that means a lot. And so that improves the trustworthiness of the data (S06).

However, our subjects agreed that provenance information plays an additional role in making a dataset trustworthy. Specifically, provenance contributed to their assessments of three constructs that are components of trustworthiness: accuracy, integrity, and authenticity. Provenance information functioned in multiple ways under Hilligoss and Rieh's framework. In some cases, provenance information acted as credibility cues, for example, direct indicators of data accuracy; other elements of provenance information served as heuristics, allowing users to indirectly assess the integrity and authenticity of data.

Accuracy

When asked what pieces of information are most interesting or important when evaluating a dataset, most subjects focused on experimental and methodological

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

parameters, such as the instrument used (S02), the characteristics of the sample used (S13), and the search algorithm (S16) or other data processing methods used (S01). These examples of provenance information help in evaluating the accuracy of data. Along with these direct indicators of data accuracy, users also relied on interactions with peripheral source cues, or their knowledge about *characteristics* of the data source. S12 noted, for example, that he could gauge how accurate data were based on what brand of instrument had been used in their generation. Others noted that the age of the instrument that produced the data is important. Data produced from an old instrument might be superseded by data from a new version (S13); the old data may not be incorrect, necessarily, but it would not be worth using, especially if newer, more accurate data were available or could be generated.

The availability of provenance information related to data accuracy was, however, a common complaint. S01, S13, S14, S15, and S17 all felt there were specific experimental parameters or methodological details that should be made available more often than they typically are. S01 noted that "the achieved mass accuracy, mass precision and false discovery rate are the most critical parameters" but are infrequently reported. S13 pointed out, however, that some of this information "may not have to be explicitly specified in the data entry forms when you're uploading data" because it is automatically encoded in raw instrument files; those parameters would only be left out if a processed file, rather than a raw file, is uploaded.

Integrity and authenticity

The integrity of data and its authenticity can be difficult to assess directly; without actual knowledge of the state of a dataset upon submission, users cannot assess for themselves whether the data have lost integrity over time or that they are what they purport to be. In order to indirectly assess data integrity and authenticity, users relied on several heuristics that draw on provenance information.

One heuristic our subjects employed is media-related. Hilligoss and Rieh defined media broadly, meaning "any media, format, or channel through which information is conveyed" (p. 1475). The two media in which subjects encounter data in the context of proteomics research are through a manuscript (either in print or online) or through a repository. Subjects indicated that published data were more credible than data that existed solely in a repository, in part because published data are more strictly controlled. For example, S15 said, "Using unpublished data […] is therefore in my view not ideal, also because Tranche permits data to be withdrawn and deleted. Information in published manuscripts is more long-lasting" (S15). Further, subjects suggested that the quality of the paper directly reflected the quality of the data; for example, if a paper did not sufficiently describe the data associated with it, S16 argued, "you probably wouldn't even bother to look further." Subjects did not, however, differentiate between media at a more granular level, either between manuscripts in different journals or between the multiple repositories available to proteomics researchers. For example, S17 commented that in terms of using specific scientific repositories to access proteomics research datasets, "[i]t doesn't matter to me if it's ProteomeCommons or PeptideAtlas or PRIDE or wherever."

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

A second heuristic was related to the source of the data. Hilligoss and Rieh found that their subjects differentiated between sources in two ways: familiar versus unfamiliar and primary versus secondary. The primary versus secondary distinction is not relevant to this study, but subjects did indicate that they would find data with associated contact information for the author more credible than data that did not include that information. As S03 noted, "If I have two equivalent data sets, right; one from one researcher and one from another researcher, and if I can contact one of the researchers and not another, I guess I'll go with the researcher I could contact." To S03, the presence of contact information serves as a source-related heuristic.

Important pieces of provenance information for our subjects were whether the data were attached to a paper that had been submitted and accepted to a journal and further, whether the author was identified and could be contacted. This information ensures an independent measure of data integrity; if the data in ProteomeCommons.org or downloaded from Tranche differ from the published description, the data cannot be trusted. It also offers a guarantee of the data's authenticity; the data are more likely to be what they purport to be if there is a visible and reachable creator. But both these pieces of information potentially cue users into another component of credibility: the expertise of the data creator.

Using provenance as an indicator of expertise

Although some subjects placed relatively less importance on knowledge of the data's author than on other factors, like experimental parameters and data quality measures (for example, S15 kindly provided an ordered list of the kinds of information important to him, with "institution and corresponding author" as the ninth entry), others noted that the identity of the data producer would play a role in their judgments about the data. S06 said that his decision to use data would be in part contingent on the reputation of the data producer: "I am less likely to look at data that comes from somebody who doesn't have a reputation that is as strong […] or they don't have a reputation that I'm aware of." S17 dissented, however, noting that reputation is not everything: "There's some really good people that put out crap data, and vice versa. There's some people that are not well known that publish great data."

Perhaps, the most critical piece of information in assessing the expertise of the data creator is the existence of a publication associated with data. Our subjects repeatedly stressed the importance of having a paper to go with data; in part, this ensures data integrity and authenticity, as described above, but additionally, a published paper ensures the expertise of the data creator because the paper has been through peer review—and in this field, reviewers assess both the publication and the data that go with it.

Using provenance together with other information

Subjects' responses to the rating exercise (see Appendix 2) indicated general approval for the provenance information specified in the MIAPE standard. Eleven subjects rated their confidence in a dataset based on the provided provenance

UNCORRECTED PROOF

Journal : Small-ext 10502 Dispatch : 27-3-2012 Pages : 21
Article No. : 9172 □ LE □ TYPESET
MS Code : ARCS185 ☑ CP ☑ DISK

Author Proof

metadata above a 4 (scale: 1–5, with 5 "completely confident" and 1 "not at all confident"), two gave themselves a 3, indicating that they were neither confident nor unconfident, and three rated themselves lower than a 3. However, their comments in response to both the rating exercise and our interview questions suggested that evaluating provenance metadata is only one part of the process they go through when determining a dataset's credibility. S17 qualified his rating, explaining that depending on what he was trying to do with the data, he would consider the metadata provided dramatically more or less useful:

> Okay, so, as far as being able to replicate the analysis that's presented, I feel that would be like a four, like pretty confident about that, right? But as far as talking about the biology, [the MIAPE provenance metadata elements] don't even, they don't say anything about the way the samples are generated, and that's three quarters of the battle. So, as far as that's concerned, that's you know, like a one.

While S17 believed, if given this information, he would know enough to replicate the analysis, the preservation metadata did not include information about processes used to generate the samples, which can dramatically affect data quality, and thus, he would not feel comfortable trusting the data without this information. S13 described "very well specified metadata" as a "minimum requirement," noting that "[metadata] wouldn't automatically make [a data set] trustworthy but […] it would certainly help evaluate it." Although provenance metadata can serve as an indicator of a dataset's trustworthiness and the expertise of its creators, even very complete metadata may not be enough. S15 further clarifies this point: "In my view, it is impossible to describe all the possible details that might have an impact on the results in a standardized manner."

As we spoke with our subjects, it became clear that they do not rely on provenance in isolation but instead drew on other available kinds of information, especially information made available in their interactions with the dataset. This squares with Hilligoss and Rieh's framework, in which users partly rely on heuristics based on metadata and the information source and also draw on what they learn by interaction with information objects and the content they carry. In particular, the peripheral information object cue most frequently identified by our subjects was the amount of metadata made available with a dataset. More metadata typically indicate higher-quality data (S11); conversely, when metadata are missing, one might "normally assume that the data is of low quality" (S01). No matter how much metadata are available, it may not be enough to make a complete judgment about a dataset's credibility.

Our subjects repeatedly expressed the importance of working with the data themselves: "You never know the quality of the data that's used in the data set until you're actually manipulating the data" (S17). S04 would have "skeptical confidence" based on complete metadata, but "would have to work with that dataset extensively to convince [him]self that it's either believable or not." S12 more colorfully explained, "My assumption is that all data sets are crap. […] Then I'll process the data set and find out if they are in fact worthwhile." This kind of "hands-on" work is key to determining the credibility of a dataset and goes hand-in-

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

hand with the evaluations enabled by provenance information. While provenance information can enable at least initial judgments about the trustworthiness of data and the expertise of the producer, some users also needed to interact directly with the data contained in the dataset to make a final decision about whether they really were credible.

Connecting data to a publication: users' reliance on the "archives of science"

Hilligoss and Rieh's framework does not easily accommodate one indicator of credibility that our subjects deemed particularly important: data's connection to a published paper. Although to some extent that interaction is captured as a media-based heuristic, our subjects indicated that their reliance on publications went deeper than that. Hilligoss and Rieh's levels of credibility assessment are embedded within the search context, which serves as a catchall for any factors that influence interactions between the levels and searcher's decisions about which indicators of credibility are most important.

S06 explained that while he felt that the provenance metadata included with the dataset are useful, ProteomeCommons.org

> is just a location for the things, and then, if you go get them, together with the paper, you should be able to figure everything out, ~~'cause~~ for me, the paper is really the guide. I want everything in the paper.

While unexpected in the context of this study, the importance researchers place on connecting data and manuscripts is not unknown. It is often framed as a response to concerns over misinterpretation or misuse of data; in this sense, connecting a paper to a dataset is an easy way to provide extensive metadata. Smit (2011) asks (hypothetically), "Is it not the official version of record, as officially peer reviewed and published, that will explain background, context, methodology and possibilities for further analysis in the best possible way, and express the intentions of the person who helped collect the data?" (n.p.). Several of our respondents echoed this opinion:

> S12: I mostly see the summary text as being valuable in helping someone who is just glancing through ProteomeCommons to figure out whether this data set is of any interest and then the citation of the paper tells them any additional information they'd want to get.
>
> S15: If I download a dataset referenced in a citation in order to use the data for my own work, I would first of all also download the original manuscript because this most likely contains additional information. […] In my view, it is impossible to describe all the possible details that might have an impact on the results in a standardized manner. […] For these reasons, I think it will always be necessary to download the original manuscript.

The manuscript can provide important additional details, even when there are good metadata associated with a dataset, and conversely, information in a paper can counteract the effect of limited metadata. For example, when S11 reviewed the

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

provenance metadata in the MIAPE standard, he first ranked his confidence in data with that metadata as a one, but when told he could imagine that there was also a paper associated with the data, he changed his tune: "Well in that case let me take a look again. […] I'd be happy to give it a four."

To some extent, the expectation of more complete information in a paper reflects a practical concern: adding metadata when depositing a dataset takes time, and even when researchers recognize the usefulness of extensive metadata, they may not provide it themselves:

> S06: Well, so here, I'm a total hypocrite, okay. So, I love it when people include a lot of information, yet I don't do it myself. And so, I include the minimum that I need to include to get the hash record… so [laughs] I can submit the darn paper, okay.

Because data tend to be well described in papers, linking a dataset to the paper that describes it solves a problem of missing or less-than-complete metadata without requiring researchers to duplicate the effort they already put into writing the methods section of the paper. For S05, the paper and the metadata in ProteomeCommons.org together are important for establishing the credibility of the data, especially with respect to the process by which it was created:

> S05: we should be able to follow each data point (in this case spectrum) from its source of origin all the way through our pipelines where it may become reference data.

But can the information contained in papers be taken at face value, as simply additional metadata? Our subjects suggest not. As S04 described, papers are "necessarily incomplete" because of restrictions on length or the particular focus of a journal. Papers tell a very specific story about the data, which may or may not help anyone to reuse the data:

> S10: You know, everyone has a healthy skepticism towards any data that's out there. That's because what you report in your paper is really your interpretation of the data.
>
> S13: For the paper the goal is to sort of, the goal isn't dissemination of data and capturing of all the data one would need for all the metadata you would need to distribute the raw data. The goal of the paper is to sort of tell a story, convey a story about the particular problem that you're studying.

Manuscripts derived from data are "a distillation and stylized version of the processes" of science (Shankar 2007, p. 1458), rather than strictly reporting the happenings of the laboratory and its occupants. They "tell the story of an ideal past in which all the protocols were duly followed" (Bowker 2005, p. 7). Papers are contingent, and thus, they are not an unproblematic record of the process by which a dataset was generated and analyzed. While on the one hand, papers can make up for a lack of metadata accompanying a dataset, on the other, they also fail to provide the complete story of the data. Neither the metadata associated with a dataset nor the paper alone is enough to establish the trustworthiness of a dataset; the provenance of the data is not established solely through metadata, but rather is instantiated through

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | □ LE | □ TYPESET
MS Code : ARCS185 | ☑ CP | ☑ DISK

Author Proof

the complex of the data and the paper associated with them. This explains our subjects' ambivalence about any individual element of provenance information. No one element is necessarily more important than another because the sum of all indicators is greater than any one part.

Discussion

This study offers a concrete example of how context, in the form of disciplinary norms for establishing trustworthiness, influences the assessment of credibility. In particular, our subjects made it clear that a norm in proteomics research is that the manuscript is the definitive source for scientific knowledge; data are secondary in importance to the manuscript. Even in the case where a researcher is looking to reuse data for her own purposes, she would turn first to the paper in which the data were published before exploring the dataset itself (S04). The credibility assessment described by Hilligoss and Rieh and explored in some detail above is contingent in this instance on the existence of a manuscript.

Additionally, this norm influences not only our subjects' interpretation of the metadata and other information cues presented to them but also their expectations of how their own work will be received. Another important part of the context in which our subjects carry out their credibility assessments is that they are not only consumers of the information available in ProteomeCommons.org, they are also information providers. Thus, there is a feedback relationship between the information our subjects used in their credibility assessments and the kind of information they provide to ensure that other researchers will find their work credible. S04 further explained that he would "very much assume" that anyone looking at his data would have already read the associated paper. S11 and S17 noted that they shared this expectation, and because of it, they often provided only limited metadata about their data in ProteomeCommons.org; anyone who wanted to use the data would get everything they needed (they hoped) from the manuscripts. The assumption of a connection to a published manuscript as a necessary component of credibility thus underlies every action these subjects take and the decisions they make within ProteomeCommons.org.

As our understanding of the nature of scientific records and the diverse forms those records can take grows, new avenues for exploring how those records are constructed, used, and related to each other have opened. For our subjects, access to one record—a dataset—and its associated metadata is not enough to understand that record's provenance; instead, they relied on what is essentially "a scientific archive." We suggest that an important next step is to move beyond a myopic focus on records and consider the archives of science: the interrelationships between the records—be they specimens, laboratory notebooks, or publications—that together participate in the production of scientific knowledge. Data provenance is meaningfully communicated to users not solely through metadata but through the arrangement of a set of records, in this case datasets and associated manuscripts.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

The importance of connecting the multiple records generated in the course of an activity is well known to archivists. By creating order and destroying disorder, archives define and create value, and in doing so, "embody a social vocation to create a special place in which a certain order of values prevail" (Brothman 1991, p. 82). Cook argues that key to meaningfully creating order—putting together an archival fonds—is the idea of creatorship: "It is impossible, therefore, to conceive, let alone identify, a fonds without having a clear understanding of the nature, scope and authority of the *creator* of the records involved and of the *records-creating process*" (1993, p. 27, emphasis original). Duranti further emphasizes that it is the relationship (what she calls archival bond) among records in a fond that gives meaning to the individual records, different than context, "[t]he archival bond is expression of the development of the activity in which the document participates, rather than of the act that the document embodies (e.g., appointment, grant, request), because *it contains within itself the direction of the cause-effect relationship*" (1997, p. 217, emphasis original).

This is the understanding that archivists have to offer in the context of scientific data repositories. Metadata standards are not enough to effectively enable reuse of scientific data; the subjects in this study explicitly needed access to the archive of records produced along with a dataset, not only the dataset itself. However, because the various records of science are often dispersed—journal articles held by publishers, laboratory notebooks on shelves and hard drives in laboratories, and data in repositories—there is a significant challenge to be overcome in linking these various items together. In ProteomeCommons.org, for example, the simple hyperlink on the study description page does not seem to be sufficient. Although the link is there, it is easy to overlook. S06 described fielding questions from readers who wanted to reuse data he had made available:

> So, you know, in the paper, if you just read the darn paper, it says this, it says it in plain language, this data is available as supplementary material on the [Proceedings of the National Academy of Sciences] website and sure enough it's there. So, you know, I can't make people find it.

S06 might not be able to make people find things, but perhaps archivists can. The current design of data repositories falls well short of Cook's vision for the future, in which "virtual 'archives without walls' [exist] on the Internet to facilitate access by the public to thousands of interlinked record-keeping systems" (2001, p. 24). How can the connection between paper and data be made meaningful and obvious to readers?

Conclusions

In this study, we find that provenance metadata, while an important part of establishing the credibility of a dataset, are not wholly sufficient for that purpose. Rather, researchers rely on other kinds of information, including their own prior knowledge. Understanding how users interact with these types of information is a

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

crucial part of enabling reuse; simply ensuring that metadata in a data repository are complete will not, in and of itself, promote data reuse. This work also speaks to the importance of understanding disciplinary norms for research, credibility, and crediting work. For proteomics researchers, it is critically important to link data with the paper it was first described in. This linkage may be more or less important for other fields, and repositories must understand what is important to their designated community in order to provide data that are seen as trustworthy.

Future comparative work in the area will be valuable in establishing similarities and differences across fields in the kinds of information users rely on to determine credibility. This study was limited in the number of interviews that could be carried out, in part because proteomics is a small and relatively new field. A larger-scale study, either in proteomics or in another discipline, would be important to validating and confirming what we have discussed here.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant No. ~~[GRANT NUMBER]~~. The authors would like to thank ~~[NAMES]~~ for their guidance on the development of this project, ~~[NAMES]~~ for their feedback on earlier drafts of this paper, as well as Philip Andrews and the staff of ProteomeCommons.org for their help and support.

Appendix 1

(1) For this first question, I'll ask you to refer to one of the documents I sent—example.pdf. When researchers write a paper and make the data accessible, they often provide a citation with some basic information about the dataset. This example shows one way of acknowledging a dataset the author used. In this case, it's longish description, and they've included information like the data producer's names, experimental parameters and the hash code you could use to find the dataset in Tranche. But this is just one example of how to do it. What kinds of information would you include if you were providing a citation for a dataset? Is that information readily available? Is it typically available if you are using data generated outside of your own lab?
(2) When you read a paper and there is a data citation, what information do you typically look for? Is there information you would like to see that is not typically included in citations?
(3) Imagine you've read a paper that cites a dataset that would be of interest in your own research, and you're considering downloading it and using it yourself. Is the information in the citation sufficient for you to decide whether to use it? If not, what other information would you need? Where would you find that information?
(4) When you contribute data to ProteomeCommons, do you include all the information that you would include in a data citation in a paper? Do you provide more information, or less? What kinds of information do you include, and what do you leave out? Do you include all the information that someone else would need to cite your dataset appropriately?
(5) Does this information help you gauge the trustworthiness of a dataset? If not, what other information would you need?

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | LE | TYPESET
MS Code : ARCS185 | CP | DISK

Author Proof

Appendix 2

Below, you'll see a list of different kinds of provenance information. Imagine you have found a dataset that you know has content that is interesting to you, and you're deciding whether to use it or not.

Element name	Description
Date stamp	The date on which the work described was initiated; given in the standard "YYYY-MM-DD" format (with hyphens).
Responsible person or institutional role	The (stable) primary contact person for this dataset; this could be the experimenter, laboratory head, line manager, etc.
Data transformation techniques	Include algorithms used, preparation or processing techniques, normalization techniques.
Analysis tools	Include software name and version, initial input parameters.
Data generation	Includes location of raw data, databases queried or specifications of equipment and conditions under which data were produced.

On a scale of 1–5, with 1 being "Not at all confident" and 5 being "Completely confident," how confident are you making a decision whether or not to use the data based only on the information in front of you?

What other information would you need to make you completely confident?

Is there any information on this list that you feel is not important or could be left out?

Appendix 3: Demographic survey

Please tell us about yourself.

I. Background

(1) Are you a: (Please shade the appropriate circle)

Graduate student

Post-doctoral fellow/researcher

Faculty member

Please specify your rank:

Lab technician

UNCORRECTED PROOF

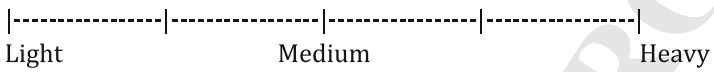
Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | □ LE | □ TYPESET
MS Code : ARCS185 | ☑ CP | ☑ DISK

Author Proof

(2) How long have you been at your current institution?
Less than 1 year
1–5 years
5–10 years
10 year or more

II. Your Experience Using ProteomeCommons

1) How heavy or light a user of ProteomeCommons are you? The scale below ranges from 'Light' to 'Heavy.' Mark the point on the scale which best matches your activity level.



(2) Why did you begin using ProteomeCommons?
[open ended]
(3) Have you or a project you've worked on ever contributed data to ProteomeCommons?
No
Yes
(4) When you contribute data, do you upload it yourself, or does someone else do it?
I upload data myself.
Another colleague is responsible for uploading data to ProteomeCommons
(5) Have you ever used data from ProteomeCommons in your own research?
I have used an entire dataset from ProteomeCommons.
I have used part of a dataset from ProteomeCommons.
I have never used data from ProeteomeCommons.
(6) Have you ever downloaded data from ProteomeCommons?
No
Yes
Another colleague is responsible for downloading data from ProteomeCommons

References

Bazeley P (2007) Qualitative data analysis with NVivo. Sage, Los Angeles
Bearman DA, Lytle RH (1985) The power of the principle of provenance. Archivaria 21:14–27. http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/11231. Accessed 28 July 2011
Bertino E, Dai C, Kantarcioglu M (2009) The challenge of assuring data trustworthiness. In: Proceedings of the 14th international conference on database systems for advanced applications. doi: 10.1007/978-3-642-00887-0_2
Bose R, Frew J (2005) Lineage retrieval for scientific data processing: a survey. ACM Comput Surv 37(1):1–28. doi:10.1145/1057977.1057978

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | □ LE | □ TYPESET
MS Code : ARCS185 | ☑ CP | ☑ DISK

Author Proof

Bowers S, McPhillips T, Ludäscher B, Cohen S, Davidson SB (2006) A model for user-oriented data provenance in pipelined scientific workflows. In: Proceedings of the international provenance and annotation workshop. http://repository.upenn.edu/cis_papers/290/. Accessed 28 July 2011

Bowker GC (2005) Memory practices in the sciences. Inside technology. MIT Press, Cambridge, MA

Brothman B (1991) Orders of value: probing the theoretical terms of archival practice. Archivaria 32:78–100. http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/11761. Accessed 28 July 2011

Buneman P, Khanna S, Tan WC (2001) Why and Where: A Characterization of Data Provenance. In: Proceedings of the 8th international conference on database theory. http://portal.acm.org/citation.cfm?id=656274. Accessed 28 July 2011

Buneman P, Chapman A, Cheney J (2006) Provenance management in curated databases. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data. doi:10.1145/1142473.1142534

Caplan P (2009) Understanding PREMIS. www.loc.gov/standards/premis/understanding-premis.pdf. Accessed 28 July, 2011

Cook T (1993) The concept of the archival fonds: theory, description, and provenance in the post-custodial era. Archivaria 35:24–37. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11882/12835. Accessed 28 July 2011

Cook T (2001) Archival science and postmodernism: new formulations for old concepts. Arch Sci 1:3–24. doi:10.1007/BF02435636

Corti L (2007) Re-using archived qualitative data—where, how and why? Arch Sci 7:37–54. doi:10.1007/s10502-006-9038-y

Dai C, Lin D, Bertino E, Kantarcioglu M (2008) An approach to evaluate data trustworthiness based on data provenance. In: Jonker W, Petković M (eds) Secure data management. Lecture notes in computer science 5159:82–89. doi: 10.1007/978-3-540-85259-9_6

Duranti L (1997) The archival bond. Arch Mus Inform 11:213–218. doi:10.1023/A:1009025127463

Duranti L (2001) The impact of digital technology on archival science. Arch Sci 1:39–55. doi:10.1007/BF02435638

Fogg B (2002) Persuasive technology: using computers to change what we think and do. Morgan Kaufmann, Amsterdam

Frey JG, Hughes GV, Mills HR, Schraefel MC, Smith GM, De Roure D (2004) Less is more: lightweight ontologies and user interfaces for smart labs. In: Proceedings of the UK e-Science all hands meeting. http://eprints.ecs.soton.ac.uk/10100/. Accessed 28 July 2011

Greenwood M, Goble C, Stevens R, Zhao J, Addis M, Marvin D, Moreau L et al (2003) Proceedings of the UK e-Science All Hands Meeting. doi:10.1.1.10.3526

Heinis T, Alonso G (2008) Efficient lineage tracking for scientific workflows. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data. doi: 10.1145/1376616.1376716

Hilligoss B, Rieh SY (2008) Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. Inform Process Manag 44(4):1467–1484. doi:10.1016/j.ipm.2007.10.001

Lauriault T, Craig B, Taylor D, Pulsifer P (2008) Today's data are part of tomorrow's research: Archival issues in the sciences. Archivaria 64:123–179. http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13156. Accessed 28 July 2011

PREMIS Editorial Committee (2008) PREMIS data dictionary for preservation metadata version 2.0. Library of Congress, Washington, DC. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf. Accessed 28 July 2011

Rieh SY (2002) Judgment of information quality and cognitive authority in the Web. J Am Soc Inform Sci Technol 53(2):145. doi:10.1002/asi.10017.abs

Rodriguez H, Andrews P, Kinsinger C (2010) Share the (Proteomics) data. Bio-IT World, (September–October 2010). http://www.bio-itworld.com/2010/issues/sept-oct/proteomics.html. Accessed 28 July 2011

Schraefel MC, Hughes G, Mills H, Smith G, Frey J (2004a) Making tea: iterative design through analogy. In: Proceedings of the 5th conference on designing interactive systems: processes, practices, methods, and techniques. doi:10.1145/1013115.1013124

Schraefel MC, Hughes GV, Mills HR, Smith G, Payne TR, Frey, J (2004b) Breaking the book: translating the chemistry lab book into a pervasive computing lab environment. In: Proceedings of the SIGCHI conference on Human factors in computing systems. doi:10.1145/985692.985696

UNCORRECTED PROOF

Journal : Small-ext 10502	Dispatch : 27-3-2012	Pages : 21
Article No. : 9172	□ LE	□ TYPESET
MS Code : ARCS185	☑ CP	☑ DISK

Author Proof

Shankar K (2007) Order from chaos: the poetics and pragmatics of scientific recordkeeping. J Am Soc Inform Sci Technol 58(10):1457–1466. doi:10.1002/asi.20625

Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. ACM SIGMOD Rec 34(3):31. doi:10.1145/1084805.1084812

Smit E (2011) Abelard and Héloise: why data and publications belong together. D-Lib Mag 17(1/2). doi:10.1045/january2011-smit

Society of American Archivists (2004) Describing archives: a content standard. Society of American Archivists, Chicago, IL

Taylor CF, Paton NW, Lilley KS et al (2007) The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25:887–893. doi:10.1038/nbt1329

Van House NA (2002) Digital libraries and practices of trust: networked biodiversity information. Soc Epistemol 16(1):99–114. doi:10.1080/02691720210132833

Vardigan M, Whiteman C (2007) ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. Arch Sci 7:73–87. doi:10.1007/s10502-006-9037-z

Zimmerman AS (2008) New knowledge from old data: the role of standards in the sharing and reuse of ecological data. Sci Technol Human Values 33(5):631–652. doi:10.1177/0162243907306704

Author Biographies

Kathleen Fear is a third-year doctoral student at the University of Michigan School of Information. She studies digital preservation of scientific and medical research data, focusing in particular on the definition of significant properties for data reuse. Prior to entering the doctoral program, Kathleen completed an MSI in preservation at Michigan and a BS is Physics from Yale University. Her advisor is Elizabeth Yakel.

Devan Ray Donaldson is a Doctoral Candidate in the School of Information at the University of Michigan, Ann Arbor. In the broad research area of digital preservation, Donaldson is interested in preservation metadata, end users of trusted digital archives/repositories, and issues related to trust in digital information. Recent publications include "Implementing PREMIS: A Case Study of the Florida Digital Archive" with Dr. Paul Conway in Library Hi Tech and "Provenance, End-User Trust, and Reuse: An Empirical Investigation" with Kathleen Fear in the proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 2011). Donaldson has been a Bill and Melinda Gates Millennium Scholar since 2002 and a Horace H. Rackham Merit Fellow since 2008.

UNCORRECTED PROOF

Journal : Small-ext 10502 | Dispatch : 27-3-2012 | Pages : 21
Article No. : 9172 | □ LE | □ TYPESET
MS Code : ARCS185 | ☑ CP | ☑ DISK